

# 本周周报

## 解聰

## 本周工作

本周对企业数据进行 small-multiple 的可视化。主要针对不同行业与广告费，研发费以及利润之间的关联进行可视化与分析。

行业代码

分析行业和商业行为的相关性。数据中有行业代码以表示所公司企业从事的行业，行业代码由有四位数组成的：

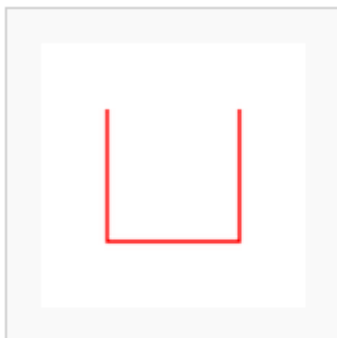
1. 数据中的行业代码前两位代表大类：  
01-09：农、林、牧、渔业，采矿业  
10-43：制造业分类  
44-46：电力、热力、燃气及水生产和供应业
2. 前两位加上后两位表示具体行业：  
比如：0152 葡萄种植；4413 核力发电。

统计了一下数据中前两位一共有 38 种可能。去掉其中两个行业后，可以考虑使用 6\*6 的矩阵来可视化这 36 种行业的相关属性

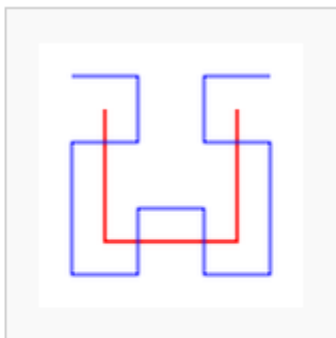
为了保证相似的行业放置在一起(比如 01-08 属于农业,应当在矩阵中放置在一起),采用 Hilbert curve 来遍历矩阵。

**Hilbert curve:**

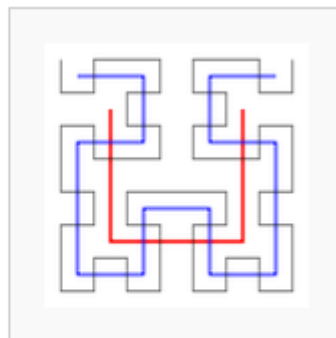
经过三次迭代后 hilbert curve 是这样的:



Hilbert curve, first  
order



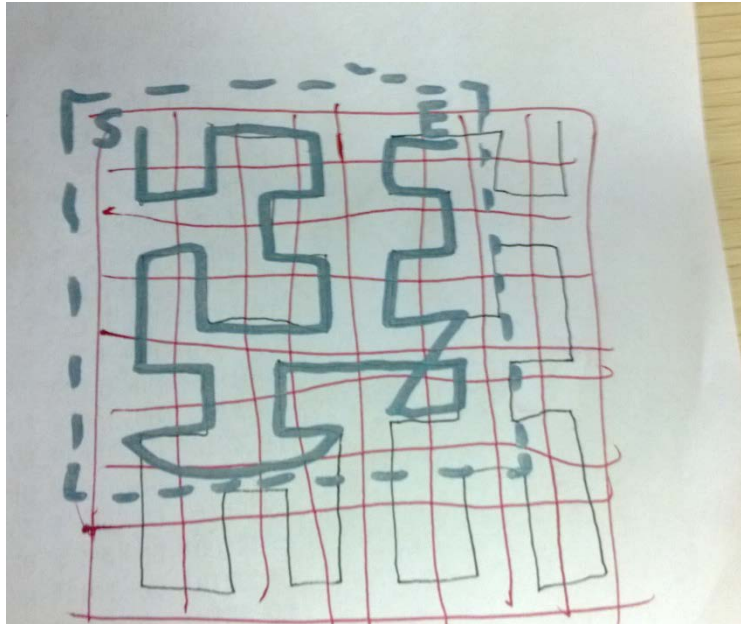
Hilbert curves, first  
and second orders



Hilbert curves, first  
to third orders

最后一张图对应  $8 \times 8$  的矩阵的遍历。

对于  $6 \times 6$  的矩阵，我们可以从  $8 \times 8$  的矩阵截取出  $6 \times 6$ ，如虚线框所示（这里截取了左上角的  $6 \times 6$  矩阵）。因此最后的遍历顺序是这样的：



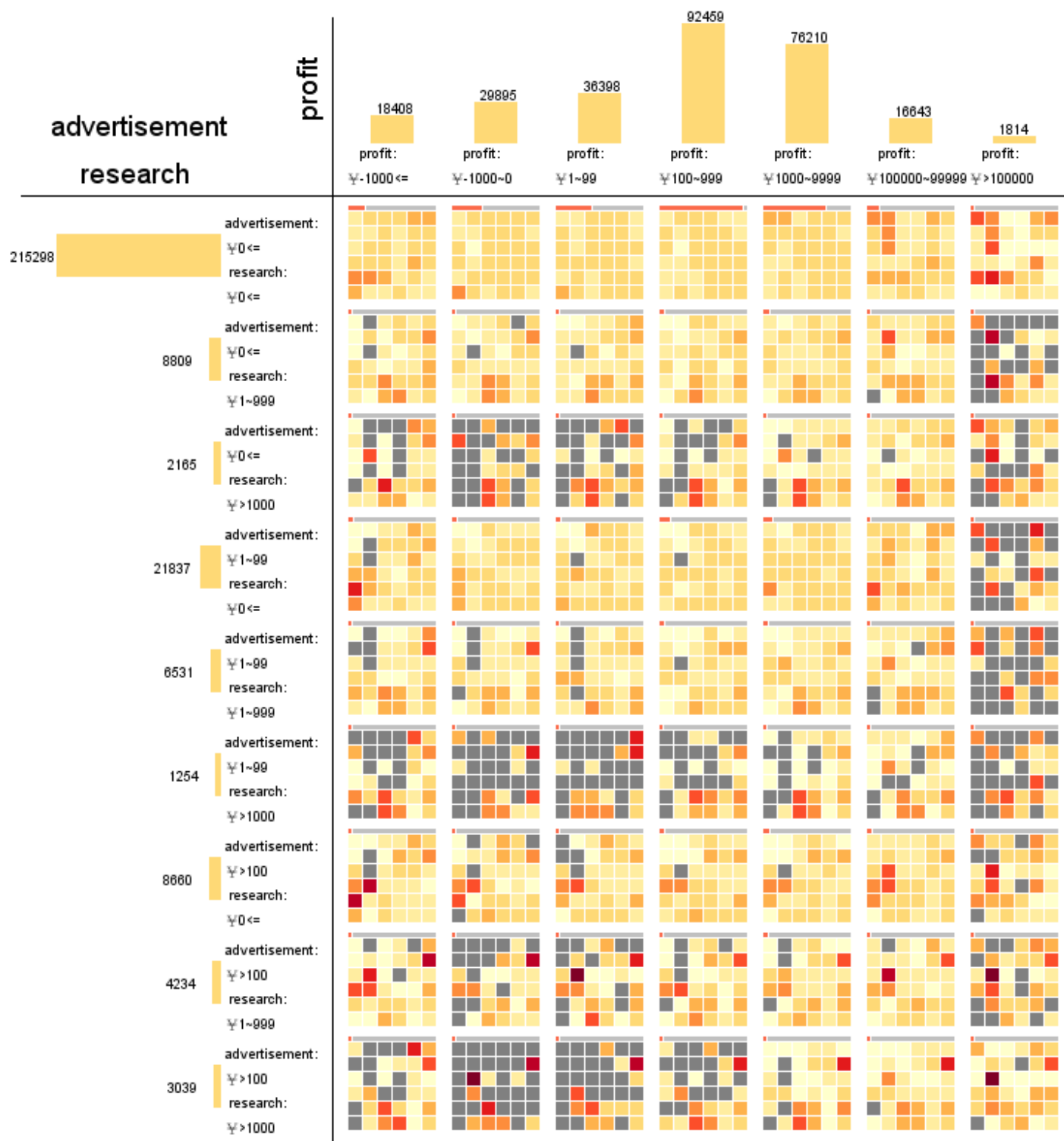
S 代表遍历起始点，E 代表遍历终点。

因此，最后行业的矩阵布局如下所示：

06 煤炭开采和洗选业	10 非金属矿采选业	13 农副食品加工业	14 食品制造业	45 燃气生产和供应业	46 水的生产和供应业
07 石油和天然气开采业	11 其他采矿业				
08 黑色金属矿采选业	09 有色金属矿采选业	16 烟草制品业	15 饮料制造业	44 电力、热力的生产和供应业	42 工艺品及其他制造业
					43 废弃资源和废旧材料回收加工业
23 印刷业和记录媒介的复制	22 造纸及纸制品业	17 纺织业	18 纺织服装、鞋帽制造业	40 通信设备、计算机及其他电子设备制造业	41 仪器仪表及文化、办公用机械制造业
24 文教体育用品制造业	21 家具制造业	20 木材加工及木、竹、藤、棕、草制品业	19 皮革、毛皮、羽毛(绒)及其制品业	39 电气机械及器材制造业	37 交通运输设备制造业
25 石油加工、炼焦及核燃料加工业	26 化学原料及化学制品制造业	31 非金属矿物制品业	32 黑色金属冶炼及压延加工业	33 有色金属冶炼及压延加工业	34 金属制品业
28 化学纤维制造业	27 医药制造业	30 塑料制品业	29 橡胶制品业	36 专用设备制造业	35 通用设备制造业

### 可视化

对利润，广告投入与研发投入进行分析：



使用颜色编码对应企业的数目，颜色越深代表企业数越多。灰色代表企业数目为 0。

## 分析

### 单个标签比较

横向对比利润发现利润较高的企业来自：

06 煤炭开采和洗选业，07 石油和天然气开采业，39 电气机械及器材制造业，45 燃气生产和供应业  
主要是能源企业。

最不赚钱的企业：

28 化学纤维制造业

研发投入较多的公司来自：

31 非金属矿物制品业，32 黑色金属冶炼及压延加工业，30 塑料制品业，29 橡胶制品业，42 工艺品及其他制造业，43 废弃资源和废旧材料回收加工业  
主要是矿产业。

**研发投入最少的产业：**

15 饮料制造业。

**广告费投入较多的是：**

24 文教体育用品制造业，26 化学原料及化学制品制造业（肥皂，洗发水，牙膏，化妆品，香精）。

**广告投入最少的企业：**

10 非金属矿采选业，11 其他采矿业，36 专用设备制造业。

**下周工作：**

1. 对企业数据进行细节分析。目前的可视化方法较适合全局概览，但是具体数据是没有办法分析的。
2. 寻找两两维度的关联。目前仅仅是从一个维度发现数据特征，下一步需要关联更多维度，比如广告费和利润的在不同行业内的关系。